The Design and Implementation of MPEG-7 Collaboration Annotation Tool (MCAT)

Wo Chang*
wchang@nist.gov

Image Group
Information Access Division
Information Technology Laboratory (ITL)
National Institute of Standards and Technology
Gaithersburg, Maryland 20899, U.S.A.

*Active member of ISO/MPEG Standards Committee and W3C SYMM WG
Keywords: MPEG-7, content description, annotation, metadata, multimedia, audiovisual

Abstract

The XML metadata technology for describing Web objects has emerged as a dominant mode of making information available for both human and machine consumption. Many online Web applications are adopting this concept for their domain-specific applications. For a metadata model to become widely used, a standardization effort is often required so that the metadata content and its structure may be interoperable between applications. The ISO MPEG standards body, with the charter of defining standard metadata to describe audiovisual content, has developed a new metadata technology called MPEG-7 content description. An effective collaboration annotation tool should also use standard metadata structures in order to improve a collaboration environment. This paper gives an overview of MPEG-7 and then presents the design and implementation of a prototype tool, MCAT. The tool supports annotation of audiovisual objects using standardized, content-description metadata so that annotations can be shared between other applications and systems.

1. Introduction

The Web has placed at our disposal a wealth of multimedia objects that must be interpreted, annotated, and indexed to be useful. The current approach for accomplishing this is based on proprietary, annotation architectures either in stand-alone environments or groupware settings. Such annotations, including synchronization information, are often the result of an iterative collaborative process leading to commonly accepted understandings among collaborators. This process can be cumbersome and time consuming even when automated tools are available. The utility of collaborations can be increased if the work is available to other applications. However, experience shows that new releases of a product may fail to be compatible even with the preceding version. To avoid this, users rely on standard formats to transfer information between versions and other applications.

The Web and related technologies provide the means for developing robust and inexpensive applications that reliably support interoperability through the use of widely accepted, non-proprietary, metadata standards. The International Organization for Standardization (ISO), Moving Picture Experts Group (MPEG) standard metadata MPEG-7 [1], a content description technology, can be an adaptable solution when dealing with audiovisual, collaborative, annotation applications. It is an eXtensible Markup Language (XML)-based technology that can be integrated easily with a Web-based annotation collaboration tool. This allows the deployment of platform-independent annotation tools and sharing of standard annotation metadata between collaborators.

This paper presents a prototype of the MPEG-7-based collaborative, annotation-tool, MCAT. The outline of this paper is as follows: Section 2 presents the core XML-based MPEG-7 content description technology, Section 3 provides the overview of the MCAT System Architecture, Section 4 maps annotation objects with MPEG-7 technology, Section 5 presents the design and implementation approach of MCAT, Section 6 discusses universal audiovisual metadata, and Section 7 summarizes the paper and outlines future work.

2. MPEG-7 and Web Technologies

2.1 MPEG-7 Content Description

Metadata description technology has become popular over the past few years. Examples include industry-defined initiatives such as the Dublin Core [2], SMPTE Metadata Dictionary [3], EBU P/Meta [4], and TV Anytime [5]. However, these industry activities are tailored to very specific application domains. Therefore, in order to support the interoperability between the above application domains, a generic, flexible, and extensible standard, metadata framework for describing audiovisual content is needed.

The MPEG-7 standard, also known as "Multimedia Content Description Interface", aims to provide standardized, core technologies allowing description of audiovisual, data-content in multimedia environments. This standard is being designed by a range of experts including content creators, broadcasters, manufacturers, publishers, intellectual property rights managers, telecommunication service providers, academia, government, and so on, to:

- Define a rich set of standardized tools to describe audiovisual content, that at a minimum, will support the above industries' defined metadata needs
- Create optimized storage solutions; high-performance content identification; fast, accurate, personalized filtering, searching, and retrieval data structures and formats
- Enable both human users and automatic systems to process the encoded, audiovisual content-descriptions

Basically, the MPEG-7 standard community desires to standardize the metadata structure and its attribute/value definitions to describe audiovisual content, as shown in Figure 1.

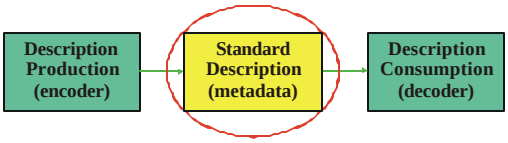

Figure 1: MPEG-7 Standard Content Description

The most challenging task of the MPEG-7 standardization effort is addressing the broad spectrum of requirements and targeted, multimedia applications. Additionally, an extensive number of important, audiovisual features from various emerging, application-domain technologies must be considered. In order to satisfy these many requirements, MPEG-7 will need to standardize common, content-description components as shown in Figure 2. These components are:

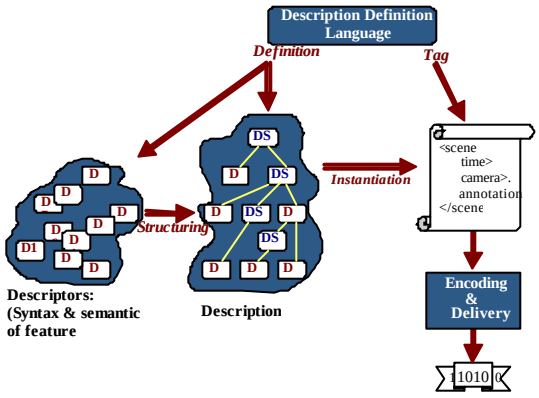

Figure 2: MPEG-7 Components
(Source: Dr. Philippe Salembier of Universitat Politecnica de Catalunya; part of MPEG-7 [1] specification))

- **Datatypes** – provides a set of description elements that are not specific to any multimedia domain. These basic datatypes are the building blocks of reusable basic types or structures employed by multiple Descriptors and Description Schemes.
- **Descriptors (D)** – defines the syntax and the semantics of each feature representation. A feature is a distinctive attribute that describes the characteristic of some object.
- **Description Schemes (DS)** – specifies the structure and semantics of the relationship(s) between components that may contain ether Ds, DSs, or both.
- **Description Definition Language (DDL)** – provides a language to create new DSs, and possibly Ds, that enables extension and modification to existing DSs.
- **Systems tools** – supports multiplexing of descriptions, or descriptions and data, synchronized multiple streams, transmission mechanisms, file format, binary encoding, and so on.

The above basic components are used to define useful functionalities within MPEG-7. These functionalities can be grouped into five major categories as shown in Figure 3. The following section describes briefly the high-level functionalities of each category:

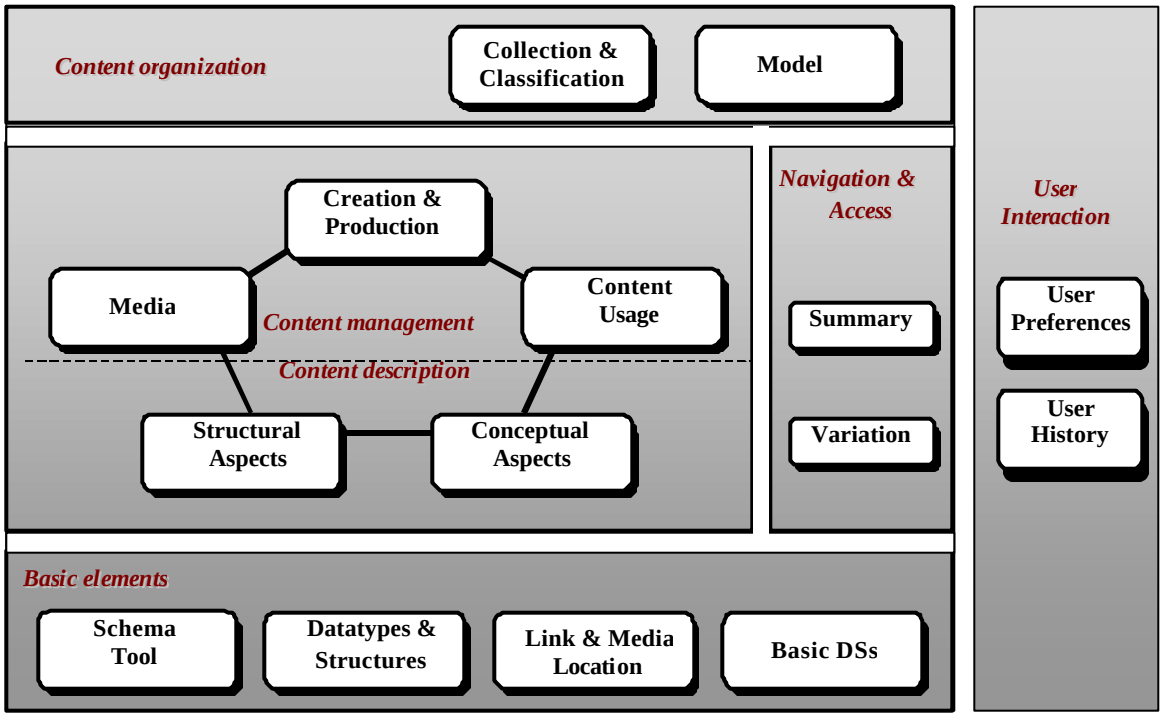

Figure 3: Overview of MPEG-7 Categories
(Source: Dr. Philippe Salembier of Universitat Politecnica de Catalunya; part of MPEG-7 [1] specification)

Basic Elements contains a group of Schema Tools for dealing with top-level root elements and packages. The Link & Media Localization group provides time, duration, and media location information. In addition, the Datatypes and Structures group offer textual annotation (either free text or structured text) and information about agent, place, graph, and so on. The Basic DSs group provides a set of predefined, description schemes.

Content Management & Description contains a Media group for handling format, coding, instances, identification, transcoding, and hints. The Creation & Production group provides title, creator, location & date, purpose, classification, and genre information. The Content Usage group covers the rights holder, access rights, usage record, and financial aspects. The Structural Aspects group provides spatial & temporal structure and elementary, semantic information. And, the Conceptual Aspects group binds relationships between events, objects, and abstract concepts.

Navigation & Access contains a Summary group for discovery, browsing, navigation, and visualization, while the Variation group handles adaptation to terminal, network, and user preferences.

Content Organization contains a Collection & Classification group for description and organization of collections of documents. The Models group provides statistical functions and structures to describe samples of audiovisual content as in a probability model.

User Interaction contains a User Preferences group for handling user identification and preferences, filtering and searching. It also provides usage history.

2.2 W3C Web Technologies

All MPEG-7 components are based on XML [6] technologies by World Wide Web Consortium (W3C) [7], these include: XML Schema [8], XPointer [9], XPath [10], DOM [11], and others. XML allows document structures to be defined, levels of subdivisions to be created, and content data to be stored and retrieved hierarchically. XML is data-centric, whereas HTML (HyperText Markup Language) is display-centric. XML focuses on how data is structured rather than on how data is displayed. XML provides mechanisms to define descriptors and description schemes that, in turn, describe audiovisual content-descriptions. MPEG-7 is based on a number of XML technologies, that are described below:

- *XML Schema* specifies the structure of instance documents and the datatype for each element and attribute. XML Schema is far more advanced than its predecessor, DTD (Document Type Definition). Improvements include: 30 more datatypes than DTD; and XML schema, but not DTD, that can extend a datatype, thereby supporting multiple elements.
- *XPointer* allows document referencing within a document while XPath deals with external document referencing. The goal of XPointer and XPath is be able to create application domain datatypes, their elements and attributes, and to be able to reference any part of the document either internally or externally.
- *DOM* provides the means for manipulating (READ and WRITE) XML structures within XML documents. SAX [12], while not a W3C standard, provides a fast, event-based parser to process (READ only) XML documents.

3. Overview of MPEG-7 Collaboration Annotation Tool (MCAT)

Collaborative annotation tools could benefit from interoperable metadata between applications [13]. MPEG-7 content-description could address this need by providing a standard, metadata framework. MCAT, an MPEG-7-based, collaborative annotation tool, demonstrates this concept. It uses the client-server model and generic HyperText Transport Protocol (HTTP) protocol to transfer audiovisual objects and user annotations to and from an HTTP server. MCAT supports spatial and temporal synchronization of audiovisual objects. Because MCAT can handle spatial and temporal synchronization, it enables collaborators to prepare and review annotations for various applications ranging from simple images, audio, or video to slide show presentations or even, multi-source synchronization applications such as in ACTS [14] and SMAT [15]. Furthermore, MCAT provides a standard, storage structure for the annotation information, so that any MPEG-7-based application can view the annotated content. Figure 4 shows the MCAT system architecture. It shows how annotations can be reviewed, annotated, and shared in a collaborative environment. The subsections that follow describe MCAT in more detail.

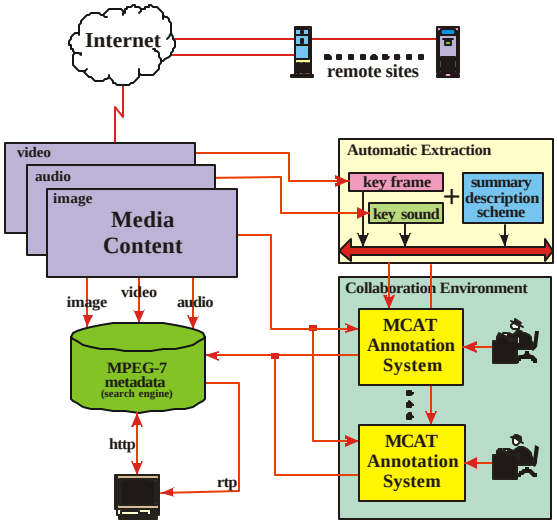

Figure 4: MCAT System Architecture

3.1 MCAT Client

Most users are familiar with platform-independent browsers. These browsers have continued to grow in functionality and have incorporated new formats as they emerge. They provide a portable, graphical user-interface. Through the use of Java, they provide the means for delivering platform-independent applications to the user. The MCAT client is a Java applet that uses browser technology. MCAT also utilizes the JMF [16] (Java Media framework) API (Application Programming Interface) to support various audiovisual objects such as WAV [17] (Waveform) and AU [18] (Audio) for audio, and QuickTime [19] and MPEG-1 [20] for video, to name a few. As JMF has matured, it has continued to support new media formats, while providing high performance at the local system.

3.2 MCAT Metadata Server

The MCAT server utilizes CGI (Common Gateway Interface) to communicate with the MCAT clients for hosting user annotations. At the user's local machine, annotations are captured and formatted using MPEG-7 predefined DSs and Ds to represent the data. An XML-based file is created and sent to the MCAT server so that the information can be shared among collaborators, as well as, between other applications and systems.

3.3 MCAT Server Authentication

It is important to have access control and user authentication for any collaborative annotation-tool so that annotated information can be protected from unauthorized use. Instead of creating our own authentication scheme, we used HTTP server authentication via CGI to communicate with MCAT clients for hosting users' annotations. The HTTP server provides two levels of authentication. At the top level, the host access level, the server can check if the incoming host has permission to access the server. At the second level, the directory access level, the server can validate the username and password to establish access rights. We can, therefore, establish user and group accounts as needed.

4. Mapping Audiovisual Annotation Objects to MPEG-7 DSs and Ds

MPEG-7 provides a rich set of description schemes (DSs) and descriptors (Ds). Currently, more than 100 DSs and 30 Ds are defined. MPEG-7 content-descriptions are comprehensive and applicable to almost any multimedia application. Therefore, MPEG-7 provides a standard framework for a multimedia-based, collaborative annotation-application.

Before mapping to MPEG-7 metadata, a list of essential metadata must be identified for any application. For a multi-user, collaborative annotation-tool, the essential metadata can be defined and categorized into the following areas:

- User Profile Metadata – information about the user (e.g. name, title, url)
- Annotation Session Metadata – session creation information (e.g. date, time)
- Annotation Metadata – actual annotation either structured (form) or non-structured (free text)
- Media Metadata – location information about actual audiovisual data (e.g. url, start time, duration)

After the metadata has been identified, it can then be mapped to MPEG-7. Table-1 shows the mapping entities and relationships from the metadata above to MPEG-7 DSs and Ds for MCAT.

	User Profile Metadata	Annotation Session Metadata	Annotation Metadata	Media Metadata
Textual DataType				
FreeTextAnnotation DataType			X	
StructureAnnotation DataType			X	
Primitive Descriptors (Ds)				
MediaLocator D				X
TemporalSegmentLocator D				X
Histogram D			X	
Media/Meta Info DS				
Person DS	X	X		
Creation Meta Info DS		X		X
Structure entities DS				
Video Segment DS				X
Still Region DS				X
Audio Segment DS				X
Semantic entities				
Event DS				X
Object DS				X
Event/Object Relation Graph				X
Summarization entities				
Summary DS			X	

Table 1: Mapping entities and relationships from MCAT to MPEG-7 DSs and Ds

4.1 MCAT Users Profile

A typical MPEG-7-based user-profile inherits XML syntax as shown in Listing 1. After parsing the metadata, MCAT can extract the proper information to construct the graphical user interface (GUI), user-profile dialog as shown in Figure 5.

```
<Name xml:lang="en">
 <GivenName>wo</GivenName>
 <FamilyName>chang</FamilyName>
</Name>
<Title>Electronic Engineer</Title>
<Affiliation>
 <Organization>
  <Name>NIST</Name>
 </Organization>
</Affiliation>
<ElectronicAddress>
 <Email>wo.chang@nist.gov</Email>
 <URL>http://smil.nist.gov/player</URL>
</ElectronicAddress>
<MediaUri>
 http://smil.nist.gov/wochang.gif
</MediaUri>
```

Listing 1: MCAT User Profile in MPEG-7 Format

Figure 5: MCAT User Profile

4.2 MCAT Annotation Session

MCAT is a multi-user, general-purpose annotation-tool used to annotate audiovisual objects. It is necessary to organize user annotations into annotation sessions. With annotation session metadata, MCAT can easily construct an annotation session directory as shown in Figure 6. Users of the system can consult this directory to locate recent changes. Basically, this directory provides general information such as date, time, author and user-added media annotations such as audio, image, video, text, and can determine the size of a given annotation instance.

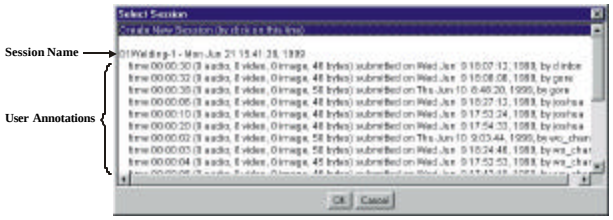

Figure 6: MCAT Annotation Session Directory Listing

4.3 MCAT Annotation Metadata

MCAT uses both MPEG-7's FreeTextAnnotation and StructuredAnnotation datatypes to capture and store the user annotations. FreeTextAnnotation is free form text and StructuredAnnotation defines the basic annotation structured in terms of *Who*, *What*, *Where*, *When*, *Why*, and *How* attributes. A typical user annotation is shown in Listing 2.

```
<FreeTextAnnotation xml:lang="en">
  Please review this video clip!
</FreeTextAnnotation>

<StructuredAnnotation>
 <Who>
  <Name xml:lang="en">ABC News</Name>
 </Who>
 <WhatObject>
  <Name xml:lang="en">Crime Site</Name>
 </WhatObject>
 <WhatAction>
   <Name xml:lang="en">
    Interview the survivors
   </Name>
 </WhatAction>
</StructuredAnnotation>
```

Listing 2: Sample of MCAT User Annotation in MPEG-7 Format

4.4 MCAT Audiovisual Metadata

For media metadata, MCAT uses MPEG-7's MediaTime and BytePosition of TemporalSegmentLocatorType and MediaUri of MediaLocator heavily since all audiovisual streams are distributed over the network. MCAT uses these descriptors to playback standalone streams, or view synchronized, multi-source streams from remote hosts. Listing 3 shows a video segment specified by the URI (Uniform Resource Identifier) of a "video.mpg" file and the relative start time with respect to the beginning of the file and the duration of the segment. The attribute provides the "demo.mpg" file with a particular byte-position offset.

```
<MediaUri>
 http://mpeg7.nist.gov/mcat/video.mpg
</MediaUri>
<MediaTime>
 <MediaRelTimePoint timeBase="MediaUri">
  PT3S
 </MediaRelTimePoint>
 <MediaDuration>PT10S</MediaDuration>
</MediaTime>

<MediaUri>
 http://mpeg7.nist.gov/mcat/demo.mpg
</MediaUri>
<BytePosition offset="1024"/>
```

Listing 3: Sample of MCAT Media Metadata in MPEG-7 Format

5. MCAT Design and Implementation Approach

The design goal for MCAT was to provide a content annotation environment between collaborators, while using a standard, annotation format for interoperability with other applications and systems. Synchronization between audiovisual objects such as audio, video, and images is important for a multimedia annotation-tool, where data needs to be analyzed and presented cohesively. MCAT provides a GUI environment that includes a navigational front-end and a tightly integrated, synchronized, multimedia presentation. Instead of relying on hard-to-control external, helper applications to handle audiovisual objects, MCAT has a built-in capability to handle images, audio, and videos. Overall, MCAT uses as many open standard components as possible. Figure 7 shows the MCAT system components.

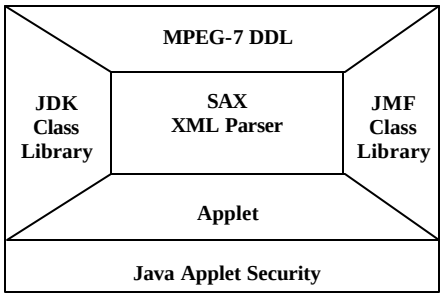

Figure 7: MCAT System Components

Specifically, MCAT uses Simple API for XML (SAX, version 2) parser for parsing; JMF for media rendering; and XML Schema 1.0 based MPEG-7 DDL for content description language.

5.1 MCAT Annotation Data and Search Engines

Even the more advanced Web-based, annotation tools available on the Internet cannot capture the intended semantics of annotated information because it is hard to index "free format" information. When searching the Internet, long lists of unwanted links are returned due to multiple denotations (e.g., 'tree' in gardens is different than the 'tree' in computer science) of keywords appearing within the indexed documents. MCAT uses the MPEG-7-based, structured, XML-metadata tags to delimit the annotation data. This captured data would enhance the search capability of any search engine, and allow more relevant search results. Moreover, MPEG-7 standard metadata goes one step further. Since all the XML tags (DSs and Ds) are predefined and standardized, MPEG-7-conformant applications or systems can intelligently perform precise searches based on the feature sets of the DSs and Ds.

Figure 8 shows the MCAT client user interface. Other than typical browser buttons/options, MCAT consists of three major areas:

- Audiovisual display area (top-left pane) – allows users to view and play audiovisual objects
- User annotation area (top-right pane) – displays user annotations with respect to the timeline
- Annotation editing area (bottom pane) – allows users to enter annotations

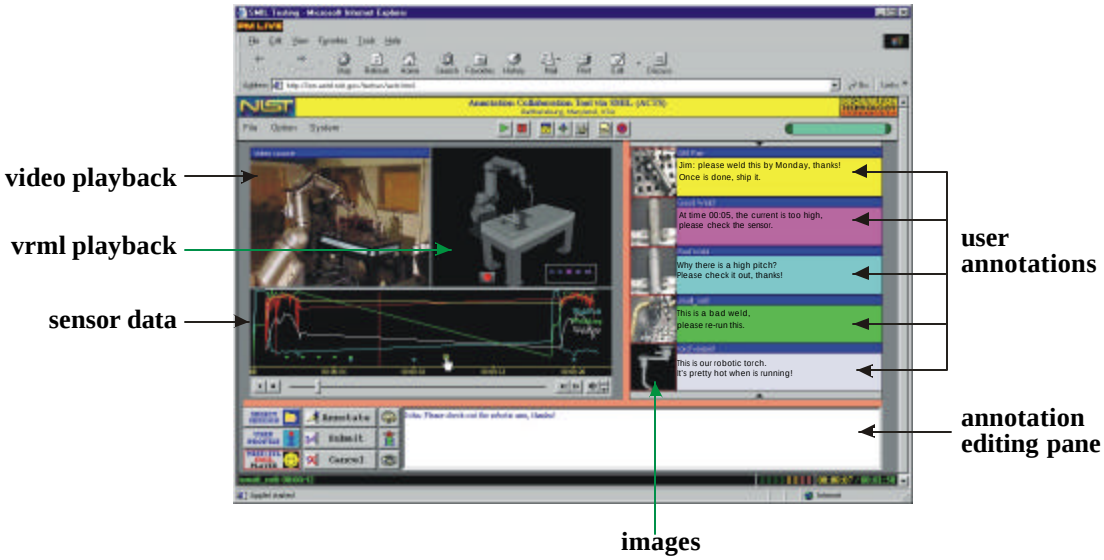

Figure 8: MCAT Client User Interface

6. Universal Audiovisual Metadata

The main goal of using standard metadata is to allow conformant applications or systems to interoperate and unambiguously exchange metadata. The success of ISO MPEG technology is based partially on its insistence on interoperability for all the standard technologies it develops. This is especially true for MPEG-7. It allows multimedia applications to be built using a standard metadata structure. Without standard metadata, companies will likely invent their own proprietary, XML structures and definitions. On the other hand, MPEG-7 has been created via the ISO, and is supported by more than 100 multimedia-based companies from over 25 countries.

6.1 Potential MPEG-7-based Applications

Multimedia applications can easily adapt and use MPEG-7 content description technology. Application areas include:

- Broadcast media selection (e.g., radio channel, TV channel)
- Digital libraries (e.g., film, image, video and radio archives)
- E-Commerce (e.g., advertising, directories of e-shops)
- Education (e.g., multimedia courses, on-line training)
- Home Entertainment (e.g., home video, game, karaoke)
- Multimedia directory services (e.g., yellow pages, tourist information)

6.2 Standard Interoperable Metadata

The main objective for ISO MPEG in developing all MPEG-related (MPEG-1/-2/-4/-7/-21) standards is to ensure that products are interoperable, either at the hardware chip-level or at the software-application level. This is true for MPEG-7, even though it covers a wide spectrum of multimedia applications. To achieve this, MPEG-7 experts are working on the topics of conformance and interoperability. MPEG-7 might use the same approach taken in MPEG-4, that defines profiles and levels as the checkpoints when dealing with different capabilities of different terminals and systems. Such a scheme is necessary because terminal configurations may vary in memory size, number of supported DSs and Ds, processing power, dynamic update on schemas, and so on.

Implementing interoperability checkpoints within MPEG-7 is a complicated issue for MPEG-7 experts, since checkpoints can be anywhere under different system configurations. However, since MPEG-7 is standardizing the content descriptions on DSs, Ds, and DDL, it is possible to interchange MPEG-7 descriptions unambiguously in textual and binary representation formats. MPEG-7 uses well-defined, standard metadata, such that the reconstructed DS/D structure is the same for all conformant, description-consuming applications and systems.

7. Conclusions and Future Work

The motivation behind building the MCAT tool is to demonstrate the advantages of using standard metadata provided by MPEG-7. This technology allows MCAT to serve collaborative tool, whose data is available to other applications. In addition, using the well-defined feature sets of DSs and Ds gives tremendous advantages for indexing and retrieval engines, because the semantics of DSs and Ds are well-defined. This paper presents the design and implementation approach of MCAT with the objectives of a) taking advantage of MPEG-7 standard metadata content description, and b) integrating standards/industry technologies (XML, SAX, JMF, Java, etc.) for developing platform-independent, portable systems.

In the future, we plan to expand the current MCAT with the following goals:

- Build an index and retrieval system based on the annotated data so that other domain applications and experts can easily analyze the data
- Investigate other content extraction tools for audio and video content
- Integrate MPEG-7 BiM (Binary Metadata format) for dynamic updates for DSs, Ds, and schema between MCAT client and server

8. Disclaimer

NIST does not endorse or recommend any of the mentioned standards, products, companies, or sites in this paper, and such mentions do not imply that the cited standards, products, companies, or sites are better or worse than similar standards, products, companies, or sites.

9. Acknowledgements

I would like to thank my colleagues at ITL for their help and encouragement, in particular, Joyce Myrick for her unceasing support and Dr. Charles Wilson, Manager of the Image Group at NIST for his support and encouragement.

10. References

[1] MPEG MDS Group, "MPEG-7 Multimedia Description Schemes WD (Version 4.1)" Doc. ISO/MPEG M6477, MPEG LaBaule, France Meeting, October 2000

[2] Dublin Core Metadata Initiative http://uk.dublincore.org/

[3] Society of Motion Picture and Television Engineers (SMPTE) http://www.smpte.org/

[4] European Broadcasting Union (EBU) http://up.ebu.ch/home_1.html

[5] TV-Anytime http://www.tv-anytime.org/

[6] T. Bray, J. Paoli, C. Sperbrg-McQueen (editors). Extensible Markup Language. http://www.w3.org/TR/REC-xml

[7] World Wide Web (W3C) http://www.w3.org

[8] Henry S. Thompson, et al (editors). XML Schema. http://www.w3.org/XML/Group/Schemas.html

[9] Steve DeRose, et al (editors). XML XPointer http://www.w3.org/TR/WD-xptr

[10] James Clark and Steve DeRose (editors). XML XPath http://www.w3.org/TR/xpath

[11] Document Object Model (DOM) http://www.w3.org/XML/Group/Schemas.html

[12] Simple API for XML (SAX, Version 2) http://www.megginson.com/SAX/

[13] Kevin K. Mills, "Introduction to the Electronic Symposium on Computer-Supported Cooperative Work", ACM Computing Surveys, Vol. 32, No.2, June 1999.

[14] Michelle Steves, Wo Chang, Amy Knutilla, "Supporting Manufacturing Process Analysis and Trouble Shooting with ACTS", Proceedings of the IEEE 8th International Workshops on Enabling Technologies: Infrastructure for Collaborative Enterprises (WETICE), June 1999.

[15] Michelle Steves, M. Ranganathan and Emile Morse, "SMAT: Synchronous Multimedia and Annotation Tool", HICSS-34 Minitrack on Collaborative Problem-Solving Environments, January 2001.

[16] JMF: Java Media Framework. http://java.sun.com/products/java/media/jmf

[17] WAV: Microsoft Waveform audio format http://www.microsoft.com

[18] AU: Sun's Audio format http://www.sun.com

[19] QuickTime: Apple Movie format http://www.apple.com

[20] Home of MPEG http://www.cselt.it/mpeg